

**TRANSMITTAL
FORM**

(to be used for all correspondence after initial filing)

TRANSMITTAL FORM (to be used for all correspondence after initial filing)	Application Number	09/825,607
	Filing Date	April 3, 2001
	First Named Inventor	Yamamoto, Yasutomo
	Art Unit	2188
	Examiner Name	Jasmine Song
Total Number of Pages in This Submission	Attorney Docket Number	16869P-025300US

ENCLOSURES (Check all that apply)

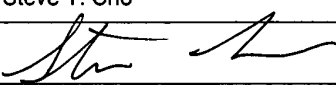
<input type="checkbox"/> Fee Transmittal Form <input type="checkbox"/> Fee Attached <input type="checkbox"/> Amendment/Reply <input type="checkbox"/> After Final <input type="checkbox"/> Affidavits/declaration(s) <input type="checkbox"/> Extension of Time Request <input type="checkbox"/> Express Abandonment Request <input type="checkbox"/> Information Disclosure Statement <input checked="" type="checkbox"/> Certified Copy of Priority Document(s) <input type="checkbox"/> Response to Missing Parts/ Incomplete Application <input type="checkbox"/> Response to Missing Parts under 37 CFR 1.52 or 1.53	<input type="checkbox"/> Drawing(s) <input type="checkbox"/> Licensing-related Papers <input type="checkbox"/> Petition <input type="checkbox"/> Petition to Convert to a Provisional Application <input type="checkbox"/> Power of Attorney, Revocation Change of Correspondence Address <input type="checkbox"/> Terminal Disclaimer <input type="checkbox"/> Request for Refund <input type="checkbox"/> CD, Number of CD(s)	<input type="checkbox"/> After Allowance Communication to Group <input type="checkbox"/> Appeal Communication to Board of Appeals and Interferences <input type="checkbox"/> Appeal Communication to Group (Appeal Notice, Brief, Reply Brief) <input type="checkbox"/> Proprietary Information <input type="checkbox"/> Status Letter <input checked="" type="checkbox"/> Other Enclosure(s) (please identify below): Return Postcard
Remarks		The Commissioner is authorized to charge any additional fees to Deposit Account 20-1430.

RECEIVED

MAR 24 2004

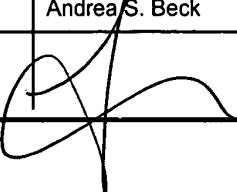
Technology Center 2100

SIGNATURE OF APPLICANT, ATTORNEY, OR AGENT

Firm or Individual	Townsend and Townsend and Crew LLP Steve Y. Cho	Reg. No. 44,612
Signature		
Date	3/17/04	

CERTIFICATE OF TRANSMISSION/MAILING

I hereby certify that this correspondence is being facsimile transmitted to the USPTO or deposited with the United States Postal Service with sufficient postage as first class mail in an envelope addressed to: Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on the date shown below.

Typed or printed name	Andrea S. Beck		
Signature		Date	3/17/04

日 本 国 特 許 庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office.

出 願 年 月 日 2 0 0 1 年 2 月 2 8 日
Date of Application:

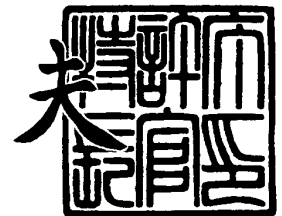
出 願 番 号 特 願 2 0 0 1 - 0 5 3 4 7 2
Application Number:
[ST. 10/C]: [J P 2 0 0 1 - 0 5 3 4 7 2]

出 願 人 株式会社日立製作所
Applicant(s):

2 0 0 4 年 2 月 1 9 日

特許庁長官
Commissioner,
Japan Patent Office

今 井 康 夫



出証番号 出証特 2 0 0 4 - 3 0 1 1 0 4 9

【書類名】 特許願

【整理番号】 K00011751A

【提出日】 平成13年 2月28日

【あて先】 特許庁長官

【国際特許分類】 G06F 17/60

【発明者】

【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所 システム開発研究所内

【氏名】 山本 康友

【発明者】

【住所又は居所】 神奈川県川崎市麻生区王禅寺 1 0 9 9 番地 株式会社日立製作所 システム開発研究所内

【氏名】 大枝 高

【発明者】

【住所又は居所】 神奈川県小田原市国府津 2 8 8 0 番地 株式会社日立製作所 ストレージシステム事業部内

【氏名】 佐藤 孝夫

【特許出願人】

【識別番号】 000005108

【氏名又は名称】 株式会社日立製作所

【代理人】

【識別番号】 100075096

【弁理士】

【氏名又は名称】 作田 康夫

【手数料の表示】

【予納台帳番号】 013088

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【プルーフの要否】 要

【書類名】 明細書

【発明の名称】 情報処理システム

【特許請求の範囲】

【請求項 1】

ホストコンピュータと、前記ホストコンピュータに接続され、複数のディスク装置を有する記憶装置とを有する情報処理システムであって、
前記ホストコンピュータは、
前記複数のディスク装置の物理的な記憶領域と論理的な記憶領域との対応関係の情報が登録される情報記憶手段と、
前記複数のディスク装置のうち第一のディスク装置に記録されたデータを第二のディスク装置に移動する場合に、移動先の論理的な記憶領域及び前記情報記憶手段から、前記第二のディスク装置の物理的な記憶領域を示す情報を取得する取得手段と、
前記取得手段によって取得された前記第二のディスク装置の物理的な記憶領域を示す情報及び移動対象となるデータが格納されている前記第一のディスク装置の物理的な記憶領域を示す情報を、前記記憶装置に転送する転送手段とを有し、
前記記憶装置は、
前記転送手段によって転送された前記情報のうち、前記第一のディスク装置の物理的な記憶領域を示す情報を使用して、移動元のデータを読み出す手段と、
前記記憶手段に記憶された前記情報のうち、前記第二のディスク装置の物理的な記憶領域を示す情報を使用して、移動先のディスク装置に複写する複写手段と、
を有することを特徴とする情報処理システム。

【請求項 2】

前記記憶装置は、
前記複写が終了したことを前記ホストコンピュータに通知する通知手段を有し、
前記ホストコンピュータは、
前記通知を受け取った後、前記情報記憶手段に登録されたディスク装置の物理的な記憶領域と論理的な記憶領域との対応関係を更新する手段を有することを特徴とする請求項 1 記載の情報処理システム。

【請求項 3】

前記転送手段は、
前記複数のディスク装置のうちの所定のディスク装置に前記情報を書き込む命令を発行する発行手段を含み、
前記記憶装置は、
前記発行手段によって発行された命令によって書き込まれた前記情報を当該記憶装置が読み出した場合に、前記読み出し手段及び前記複写手段を実行することを特徴とする請求項 1 又は 2 記載の情報処理システム。

【請求項 4】

前記記憶装置は、
前記複写手段によって複写されたデータに対してアクセスがある場合には、当該アクセスがあったことを記録するアクセス記録手段と、
前記アクセス記録手段に記録された内容に基づいて、移動元のデータと移動先のデータとのデータの内容を一致させる手段とを有することを特徴とする請求項 1、2 又は 3 記載の情報処理システム。

【請求項 5】

複数のディスク装置を有する記憶装置と接続される接続部と、
前記複数のディスク装置の物理的な記憶領域と論理的な記憶領域との対応関係の情報が登録されるメモリと、
中央演算部とを有し、
前記中央演算部は、
前記メモリから、データの移動先となる前記論理的な記憶領域に対応する第一のディスク装置の物理的な記憶領域を示す情報を取得する取得手段と、
前記取得手段によって取得された前記第一のディスク装置の物理的な記憶領域を示す情報及び前記データが格納されている第二のディスク装置の物理的な記憶領域を示す情報を、前記接続部を介して前記記憶装置に転送する転送手段と、
前記データの移動が終了した後、前記メモリに登録された前記複数のディスク装置の物理的な記憶領域と論理的な記憶領域との対応関係を更新する手段とを有することを特徴とする情報処理装置。

【請求項 6】

前記取得手段は、
前記メモリから論理的な記憶領域が割り当てられていない前記複数のディスク装置の物理的な記憶領域を検索する手段と、
前記検索する手段によって検索された前記物理的な記憶領域を前記第一のディスク装置の物理的な記憶領域として取得する手段とを含むことを特徴とする請求項 5 記載の情報処理装置。

【請求項 7】

前記転送手段は、
前記記憶装置が有する複数のディスク装置のうち、所定のディスク装置に、前記情報をデータとして書き込む命令を発行するものであることを特徴とする請求項 6 記載の情報処理装置。

【請求項 8】

ホストコンピュータに接続される接続部と、
複数の記憶領域と、
前記複数の記憶領域のうち、前記ホストコンピュータに使用されていない記憶領域の情報が登録されるメモリと、
前記複数の記憶領域へのデータの入出力を制御する制御部を有し、
前記制御部は、
前記接続部から入力される情報に基づいて、前記メモリに登録された記憶領域のうちのいずれかを選択する手段と、
前記選択された記憶領域に、他の前記記憶領域からデータを複写する複写手段とを有することを特徴とする記憶装置。

【請求項 9】

前記接続部から入力される情報とは、所定の条件を満たすことを要求する情報であり、
前記メモリには、複数の記憶領域の性質が登録され、
前記制御部は、
前記要求を満たす記憶領域を、前記メモリに登録された前記記憶装置の性質に基

づいて検索する検索手段と、前記検索された記憶領域についての情報を前記接続部から出力する出力手段を有することを特徴とする請求項 8 記載の記憶装置。

【請求項 10】

ホストコンピュータ及び前記ホストコンピュータに接続され、複数のディスク装置を有する記憶装置を有する情報処理システムにおいて、前記複数のディスク装置内でデータを再配置する方法であって、
前記ホストコンピュータにおいて、
再配置の対象となるデータが格納されている第一のディスク装置を特定し、
再配置先となる第二のディスク装置の情報を取得し、
特定した前記第一のディスク装置の情報及び取得した前記第二のディスク装置の情報を前記記憶装置に送信し、
前記記憶装置において、
送信された前記第一のディスク装置の情報に基づいて前記第一のディスク装置からデータを読み出し、
送信された前記第二のディスク装置の情報に基づいて前記第二のディスク装置に前記第一のディスク装置から読み出したデータを格納し、
前記読み出されたデータの格納が終了したら、前記ホストコンピュータに格納の完了を通知し、
前記通知の後、前記ホストコンピュータにおいて、前記複数のディスク装置と論理的な記憶領域との対応関係を記録したテーブルを変更することを特徴とするデータの再配置の方法。

【請求項 11】

ホストコンピュータ及び前記ホストコンピュータに接続され、複数のディスク装置を有する記憶装置を有する情報処理システムにおいて、前記複数のディスク装置内でデータを再配置するコンピュータプログラムであって、
前記ホストコンピュータにおいて、
再配置の対象となるデータが格納されている第一のディスク装置を特定し、
再配置先となる第二のディスク装置の情報を取得し、
特定した前記第一のディスク装置の情報及び取得した前記第二のディスク装置の

情報を前記記憶装置に送信するプログラムと、
前記記憶装置において、
送信された前記第一のディスク装置の情報に基づいて前記第一のディスク装置からデータを読み出し、
送信された前記第二のディスク装置の情報に基づいて前記第二のディスク装置に前記第一のディスク装置から読み出したデータを格納し、
前記読み出されたデータの格納が終了したら、前記ホストコンピュータに格納の完了を通知するプログラムとから構成され、
前記ホストコンピュータ側のプログラムは、前記通知の後、前記複数のディスク装置と論理的な記憶領域との対応関係を記録したテーブルを変更することを特徴とするデータの再配置を行うコンピュータプログラム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、ホストコンピュータ（ホスト）及び該ホストと接続される記憶装置を有する計算機システムに係り、特に、計算機システムが有する記憶装置に格納されたデータの移動を支援する機能に関する。

【0002】

【従来の技術】

一般に、計算機システムを構築する際には、ネットワークやディスク装置などのリソースがボトルネックとならないようにシステム設計が行われる。特に、プロセッサなどに比べて低速な外部記憶装置は、性能上のボトルネックとなりやすく、システム設計にあたって様々な対策が施される。その一つとして、記憶装置上のデータ格納形式の最適化が挙げられる。

例えば、高頻度でアクセスされるデータをより高速なディスク装置に格納したり、データを複数のディスク装置に分散配置することでデータアクセスの性能を改善できる。また、いわゆる R A I D (Redundant Array of Independent Disk) 装置を使用する場合には、R A I D 装置のデータアクセスのシーケンシャル性を考慮して、R A I D レベル(冗長構成)を決定すれば、アクセス特性に適合したデー

タの格納が可能となる。

【0 0 0 3】

システム設計という視点から見た場合には、システムにおいて使用されるデータの格納形式を考慮して、各データに割り当てられるディスク装置の容量を決定することが必要となる。具体的には、データベース(DB)におけるDB表の領域サイズや、ファイルシステム(FS)におけるFSサイズの決定がこれに相当する。通常、計算機システムの運用に伴って取り扱われるデータは増加していく。このため、システム設計時には、関連業務における過去の実績などを元にデータ量の増加率を予測し、次に保守可能な時期までに予測されるデータ量の増加に耐えられるだけの空き領域を確保するようにディスク装置の容量を決定して、それぞれのデータ領域が定義される。

【0 0 0 4】

上述のように、システム設計に際しては、データアクセス性能の向上およびデータ量の増加を考慮して、記憶装置とデータの格納形式を組み合わせることが重要となる。この組み合わせの決定を支援する手段として、Logical Volume Manager (LVM) がある。

【0 0 0 5】

LVMとは、実際のディスク装置の任意の部分領域をまとめ、論理的に一つのボリューム（これを論理ボリュームと言い、以下「LV」と称する。）としてホストに提供するものである。LVMは、LVを管理し、LVの作成、削除、サイズ変更(拡大/縮小)なども行う。LVMは、LVを多重化するミラーリング機能や、LVを小領域(ストライプ)に分割して、複数の物理ボリューム(PV)に分散して配置するストライピング機能も有する。

【0 0 0 6】

LVMを用いた場合、ユーザは、DB表やFSといったデータを格納する領域をPVではなくLV上に配置することになる。このようにすることで、データの格納形式の選択又は管理が容易となる。例えば、LV上にFSを配置することで、通常1台のディスク装置、またはその1つのパーティションに1つしか割り当てられないFSを、複数のディスク装置に分散して配置することが可能となる。

また、ファイルの容量の増加に合わせてLVを拡張することで、最小限の手間でFSを拡張(再構成)することが可能となる。

【0007】

【発明が解決しようとする課題】

計算機システムで業務運用を続けていくと、データの格納形式を見直す必要が生じることがある。見直しの契機としては、データのアクセス傾向やアクセス特性の変化、当初見積もりとは異なるデータ量の変動などといった、システム設計時に想定されていた業務モデルの見直しによるもの、ディスク装置の追加や高速なリソースへの置き換えなどといった、前提とする物理リソース構成の変更によるもの、拡張を繰り返すことによる断片化(fragmentation)したLVやDB表の連続化などといったデータ管理方式上必然的に生じるものなどが考えられる。このような場合、データの格納形式を見直し、データを再配置することで、システムの性能を改善することができる。

【0008】

しかし、記憶装置に格納されたデータの再配置を行うには、従来の技術においては、ホストの介在によるデータ転送が必要となる。

【0009】

今、複数のPVに分散されたLVを一つのPVにまとめる場合のデータ再配置処理の手順を以下に示す。

- (1) 処理対象となるLVの容量分の領域をPV上に確保する。
- (2) LVからホストがデータをリードし、それを新しいLVの領域にライトする。
- (3) (2)を繰り返して全データをコピーし終えたら、LVとPVマッピング情報を変更する。

【0010】

このように、LVを再構成するには、LV全体という大量のデータ転送が生じるため、データ転送元/先のPVに大量の入出力(I/O)が発生する。同時に、ホストおよびチャネルにも多大な負荷がかかり、他のLV内のデータを対象とした実行中の処理にも性能上悪影響を及ぼしてしまう。

【0011】

また、このようなデータ再配置処理の間は、再配置の対象となるデータへのアクセスについて、少なくともデータの更新を抑止する必要がある。例えば、LVの再構成の場合、通常、LVをオフライン(Unixではアンマウント)とし、再構成完了後にオンライン(Unixではマウント)にすることで、データ再配置中のデータのアクセスを抑止する。データアクセスの対象となるLVがオフラインとなるため、そのLVを利用する業務は、データ再配置処理の間、停止することになる。

【0012】

このため、LV再構成等のデータ再配置を行う場合には、そのデータにアクセスする業務を所定の期間中断することができる時間帯に行わなければならない。従って、計算機システムの保守作業に時間的制約が生じてしまう。

【0013】

本発明の目的は、記憶装置に格納されたデータを再配置する際に、移動元データ領域から移動先データ領域へのデータ転送によるホストおよびチャネルの負荷を軽減することにある。

【0014】

また、本発明の別の目的は、データ再配置により、データへのアクセスが不可となる期間を可能な限り削減し、そのデータを利用する業務が中断される時間を削減することにある。

【0015】**【課題を解決するための手段】**

上記目的を達成するために、本発明では、ホストコンピュータと、前記ホストコンピュータに接続され、複数のディスク装置を有する記憶装置とを有する情報処理システムで、以下の構成とする。ホストコンピュータは、複数のディスク装置の物理的な記憶領域と論理的な記憶領域との対応関係の情報が登録されたテーブルを有する。ホストコンピュータは、第一のディスク装置に記録されたデータを第二のディスク装置に移動する場合に、第二のディスク装置の物理的な記憶領域を示す情報をテーブルから取得する。ホストコンピュータは、取得した情報を、記憶装置に転送する。データの移動が終了した後、ホストコンピュータは、テ

ーブルに登録された情報を更新する。また、記憶装置は、ホストコンピュータから転送された前記情報を用いて、移動元のデータを移動先のディスク装置に複写する。

【0016】

以上の構成により、データを再配置する場合のデータの転送を記憶装置側で行うことで、データ転送によるホストおよびチャネル負荷を軽減することができる。

【0017】

また、記憶装置でのデータ転送処理を、単純に領域間コピーで実現するのではなく、転送元と転送先のデータ領域で一時的に内容の同期をとるペアを形成し、データ転送中およびデータ転送完了後に転送元データ領域に対して行われたデータ更新を全て転送先へ反映する構成とすることも出来る。

【0018】

また、ホストコンピュータではなく、記憶装置でディスク装置の仮想的な記憶領域の管理を行う構成も考えられる。

【0019】

【発明の実施の形態】

図1は、本発明が適用された計算機システムの第1実施形態における構成を示す構成図である。

【0020】

本実施形態における計算機システムは、ホスト100及び記憶装置110を有する。ホスト100と記憶装置110は、SCSIバス等の通信線で接続されている。両者間の交信は通信線を介して行われる。

【0021】

ホスト100は、CPU101、主記憶102及びチャネル103を有し、それぞれが内部バスで接続される。

【0022】

CPU101は、データベース等のアプリケーションプログラム（AP）を実行する。APが動作する際のメモリ割り当てや、記憶装置110への入出力制御

は、CPU101で動作するオペレーティングシステム（OS）およびOSに関連するソフトウェアで行われる。LVM142も、OSに関連するソフトウェアである。LVM142は、記憶装置110により提供されるPVの記憶領域をまとめてLVとして仮想的にAPに提供する。

本実施形態では、CPU101で、LVM142が実行される。LVM142は、後述するLV-PV対応情報141などの制御情報を用いてLVの制御を行う。

【0023】

主記憶102には、アプリケーションプログラム、OS、LVM等のOS関連ソフトウェアの実行オブジェクトコードや各ソフトウェアが使用するデータや制御情報等が格納される。

【0024】

図3及び図4は、LV-PV対応情報141のテーブル図である。
LV-PV対応情報141は、LVが対応するPV（又はPVが対応するLV）を示す情報である。LV-PV対応情報141のテーブルには、LV管理情報300とPV管理情報400が、1つのLV（PV）ごとに設けられる。

【0025】

LV管理情報300は、PVリスト301、LE数302、LEサイズ303及びLE-PE対応情報310のエントリを有する。
PVリスト301には、LVに対して割り当てられているPVを示す情報が登録される。LVとPVは、それぞれLE (Logical Extent) およびPE (Physical Extent) という同じサイズの小領域に分割されている。LEに対してPEを割り当てることで、LVの物理配置の自由度が向上する。LE数302には、LVに含まれるLE数が登録される。LEサイズ303には、LEのサイズを示す情報が登録される。LE-PE対応情報310のエントリは、LE番号311に割り当てられるLEに対応するPV名312およびPE番号313のエントリを有する。LE-PE対応情報310には、LEが割り当てられているPEを示す情報が登録される。

【0026】

PV管理情報400は、LV管理情報とは逆に、各PVに割り当てられるLVの情報を示すものである。

【0027】

LV管理情報400は、LVリスト401、PE数402、PEサイズ403及びPE-LE対応情報410のエントリを有する。

LVリスト401には、PVに対して割り当てられているLVを示す情報が登録される。PE数402には、PVに含まれるPE数が登録される。PEサイズ403には、PEのサイズを示す情報が登録される。PE-LE対応情報410は、PE番号411に割り当てられるPEに対応するLV名412およびLE番号413のエントリを有する。PE-LE対応情報410には、PEが割り当てられているLEを示す情報が登録される。

【0028】

主記憶102には、上記した情報以外にも、PVへアクセスするために必要な情報が保持される。例えば、PVへアクセスするための経路情報として、接続チャンネル103の番号および記憶装置110のポート114の番号や、記憶装置110内でのデバイス番号(以下、PV番号と呼ぶ)が格納される。

【0029】

チャンネル103は、通信線を介して記憶装置110への入出力の制御を行うコントローラである。チャンネル103は、通信線上の要求コマンドの送信、完了報告の通知、対象データの送受信、通信フェーズ管理等、通信プロトコルの制御を行う。通信線がSCSIバスの場合、SCSIアダプタカードが、チャンネル103に相当する。

【0030】

記憶装置110は、ホストとの接続の制御を行うポート114、ディスク装置150、記憶制御プロセッサ111、制御メモリ112及びディスクキャッシュ113を有する。記憶装置110の可用性を向上させるため、記憶装置110の各部位は多重化されている。したがって、ある部位に障害が発生した場合にも、残った正常な構成部位による縮退状態での運用が可能である。

【0031】

記憶装置 110 が、ディスク装置 150 が複数台接続された R A I D システムの場合、記憶制御プロセッサ 111 による論理物理対応管理によるエミュレーションを使用することによって、複数台のディスク装置 150 を、論理的に 1 台又は複数のディスク装置としてホスト 100 に認識させることができる。ただし、本実施形態では、説明を簡略化するため、ホスト 100 がアクセスする P V、すなわち記憶装置 110 での論理ディスク装置は、ディスク装置 150 と一対一に対応しているものと仮定する。

【0032】

記憶制御プロセッサ 111 は、ホスト 100 から P V へのアクセスを受け付け、ディスク装置 150 とディスクキャッシュ 113 間のデータ転送、ディスクキャッシュ 113 とホスト 100 間のデータ転送、記憶装置 110 内でのディスク装置 150 の論理物理対応の管理、ディスクキャッシュ 113 の領域管理などを行う。

【0033】

ディスクキャッシュ 113 には、ホスト 100 からのライトデータや、ディスク装置 150 からのリードデータが、転送先に送信されるまでの間、一時的に保持される。ディスクキャッシュ 113 に格納されたデータは、L R U (Least Recently Used) などの方式で管理される。ディスクキャッシュ 113 を用いて、ライトデータをホストの I / O 要求とは非同期にディスク装置 150 へ書き込む処理も行われる。これらのキャッシュの制御方法は従来公知の技術であるので、説明は省略する。

【0034】

制御メモリ 112 には、ディスク装置 150 への入出力を制御するために記憶制御プロセッサ 111 が用いる各種制御情報のテーブルが格納される。制御情報のテーブルには、ディスクキャッシュ 113 の領域割り当て管理に用いるキャッシュ管理情報 144、論理ディスク装置とディスク装置 150 との対応等の管理に用いられる記憶装置管理情報 143、ホスト 100 から指示された領域間のデータ転送処理の対象範囲や処理進捗状況等の管理に用いられるデータ転送領域情報 145 がある。

**【0035】**

図5は、データ転送領域情報145のテーブル図である。
転送元範囲情報501及び転送先範囲情報502には、ホスト100から指示されたデータ転送の対象となるデータ領域の範囲を示す情報が登録される。本実施形態では、転送元/先のデータ領域が不連続であるケースを想定し、転送元範囲情報501及び転送先範囲情報502には、連続する部分領域毎に、その部分領域を含むPV番号、PV内相対アドレスで示される先頭位置およびサイズを連ねたリストが登録される。尚、転送元/先のデータ領域の総容量は等しくなければならない。

【0036】

進捗ポインタ503には、データ転送処理で転送されたデータの量を示す情報が登録される。進捗ポインタ503に登録された情報を用いて、データ転送処理の進捗が管理される。同期状態504と差分ビットマップ505は、本実施形態では使用されないので、説明を省略する。

【0037】

本実施形態でのCPU101と記憶制御プロセッサ111の動作を説明する。

【0038】

ユーザ又は保守員は、各LVのPVへの割り当て状態等の情報から、特定のLVについて再構成が必要であることを判断すると、LVの再構成の指示を行う。本実施形態では、図2に示すように、lv0が、2台のPV、pv0とpv1上にlv0_0とlv0_1として格納されているとき、lv0をpv2内領域に新規に確保されたlv0_0'とlv0_1'に移動する指示を出した場合について説明する。

【0039】

LVの再構成は、CPU101で動作するデータ再配置処理131及び記憶装置110の記憶制御プロセッサ111で動作するコマンド処理132が協働することで実現される。

【0040】

図6は、CPU101が行うデータ再配置処理131のフロー図である。
データ再配置処理131は、ユーザ等がLVの再配置を指示した場合に実行され

る。データ再配置処理131では、その処理が実行される前に、再配置対象であるLV名と再配置先のPV名が取得される。

【0041】

CPU101は、処理対象となるLVに対するアクセスを抑止するために、対象となるLVをオフラインにする。LVのオフラインは、例えば、OSがUNIXである場合、デバイス(LV)をアンマウントすることで実現できる(ステップ601)。

【0042】

CPU101は、LV-PV対応情報141のLV管理情報300を参照して、対象となるLVのデータが格納されているPVおよびPEを求める。CPU101は、LE数302とLEサイズ303から、対象となるLVサイズを算出する。対象となるLVの全体もしくは一部が既に転送先PVに格納されている場合には、CPU101は、転送先PVに格納されている部分については、転送を行わない。ただし、以下の説明では、転送対象からはずされる部分は存在しないと仮定する(ステップ602)。

【0043】

CPU101は、対象となるLVの転送対象領域サイズ分のPEを、転送先PV上に確保する。具体的に、CPU101は、PV管理情報400のPE-LE対応情報410を参照して、LEに未割り当てのPEを求め、転送先として必要となる容量分のPEを確保する。PEの確保は、転送先PVおよびPEが記憶されるだけでもよい。しかし、他の処理によりPEを別の目的で確保されてしまう可能性がある場合は、転送に使用するPEをあらかじめ確保しておく。具体的には、PVのPE割り当ての変更を一定期間禁止することを示す構成変更処理排他フラグをPV毎に設ける、構成変更処理排他フラグをPV毎ではなくPE毎に設ける、又はデータ転送が完了する前に、PE-LE対応情報410の対象PE項目を、転送元LVに割り当て済みであるように変更する等の方法が考えられる(ステップ603)。

【0044】

転送先のPVを確保したら、CPU101は、転送元のPV領域をいくつかの

部分領域に分割し、各部分領域毎のデータ転送処理要求を記憶装置 110 に発行する。記憶装置 110 へのデータ転送処理の要求は、既存のプロトコルで準備されている標準の入出力コマンドではなく、データ転送処理の要求用に新たに追加した専用のコマンドを用いて行われる。P V 領域の分割は、適当なサイズで転送元領域を先頭から順に区切って行われる。分割に適当なサイズは、記憶装置 110 へ要求されたデータの転送処理に要する時間と、要求元ホスト 100 でのその要求に対する応答時間として許容可能な時間とから決定される。転送処理要求のコマンドが P V 毎に発行されるので、転送元 L V が複数の P V 上に分散して配置されている場合には、部分領域が 2 つの P V にまたがらないように分割する必要がある。転送処理要求のコマンドには、転送元 P V の先頭アドレスとサイズ、および転送先の P V 番号、P V 内先頭アドレス、データサイズが含まれる。チャンネル 103 を介してデータ転送要求コマンドを記憶装置 110 へ送出了ら、CPU 101 は、記憶装置 110 からの完了報告を待つ（ステップ 604）。

【0045】

ステップ 604 で送信したデータ転送処理要求コマンドの完了報告を受信した CPU 101 は、転送元 L V について全対象領域のデータ転送が完了したかをチェックする。データ転送が完了していない部分領域があれば、CPU 101 は、ステップ 604 の処理に戻る（ステップ 605）。

【0046】

対象領域の全てのデータが転送されたら、CPU 101 は、転送の対象となった L V が転送先 P V へ対応づけられるよう、L V - P V 対応情報 141 を変更する。具体的には、CPU 101 は、L V 管理情報 300 の P V リスト 301 に登録された情報を転送先 P V を示す情報に変更し、L E - P E 対応情報 310 の P V 名 312 と P E 番号 313 に登録された情報を、転送先 P V の P E を示す情報となるよう変更する。CPU 101 は、転送先 P V の P V 管理情報 400 の L V リスト 401 に、転送対象となった L V を追加し、P E - L E 対応情報 410 の L V 名 412 と L E 413 番号を、転送先 L V の L E に対応するよう変更する。CPU 101 は、転送元 P V の P V 管理情報 400 の L V リスト 401 から転送対象となった L V を削除し、P E - L E 対応情報 410 の転送元 P E 項目を L V

未割り当てに変更する（ステップ606）。

【0047】

その後、CPU101は、転送対象のLVのオフラインを解除し、処理を終了する（ステップ607）。

【0048】

図7は、記憶装置110で実行される、コマンド処理132のフロー図である。コマンド処理132は、ホスト100のコマンドを記憶装置110が受領した時に実行される。

【0049】

記憶装置110は、ホスト100からディスク装置150に発行された処理要求コマンドの種別をチェックする（ステップ701）。

【0050】

コマンドがデータ転送処理要求であった場合、記憶装置110は、コピー処理133を起動し、その完了を待つ（ステップ702）。

【0051】

コマンドがリード要求であった場合、記憶装置110は、そのリード要求に対して、対象となるデータがディスクキャッシュ113上に有るか否かをチェックする。必要ならば、記憶装置110は、キャッシュ領域を確保して、ディスク装置150から対象となるデータを、確保したキャッシュ領域へ読み出し、ホスト100へ転送する（ステップ703～ステップ707）。

【0052】

コマンドがライト要求であった場合、記憶装置110は、ディスクキャッシュ113のキャッシュ領域を確保し、ホスト100から受け取ったデータを一旦そのキャッシュ領域に書き込む。その後、ディスク装置150へライトする（ステップ709～ステップ712）。

【0053】

記憶装置110は、要求処理の完了をホスト100へ報告し、処理を終了する（ステップ708）。

【0054】

図8は、記憶装置110が行うコピー処理133のフロー図である。コピー処理133は、記憶装置110がデータ転送要求のコマンドを受けとったときに実行される。

【0055】

データ転送要求を受け取った記憶装置110は、そのデータ転送要求について転送元/先データ領域の指定情報の妥当性をチェックする。

具体的には、転送元/先データ領域のサイズが同じであるか、既に別のデータ転送要求の転送元/先データ領域として設定されていないか等をチェックする。指定情報が妥当でない場合、記憶装置110は、エラーをホスト100へ報告する(ステップ801)。

【0056】

エラーが発見されない場合、記憶装置110は、データ転送処理要求に対応するデータ転送領域情報145を格納する領域を制御メモリ112上に割り当て、初期化する。具体的には、転送元範囲情報501と転送先範囲情報502を、受け取ったデータ転送処理要求に含まれる情報を元に設定し、進捗ポインタ503を初期値である0に設定する(ステップ802)。

【0057】

設定が済んだら、記憶装置110は、転送対象となる転送元ディスク装置150のデータ領域の先頭から、順次データをディスクキャッシュ113にリードし、リードしたデータを転送先ディスク装置150へライトする。このときの一回のデータ転送のデータ量は、ディスク装置150におけるヘッド位置づけオーバーヘッドを考慮すると、ある程度大きいサイズのデータ領域であることが望ましい。しかし、1回のデータ量が大きすぎると、ディスク装置150と同じバスに接続される他のディスク装置150に格納されたデータにアクセスする他の処理に悪影響を及ぼす可能性がある。したがって、一回のデータ転送のデータ量は、コピー処理133に期待される処理速度と、他の処理への影響とを考慮して決定する必要がある(ステップ803～804)。

【0058】

ステップ804で転送先への書き込みを終えたら、記憶装置110は、転送し

たデータ容量に応じて、進捗ポインタ 503 の内容を更新する。

【0059】

記憶装置 110 は、進捗ポインタ 503 を参照して、全対象データについてコピー処理が完了したかをチェックする。コピー処理が完了していない部分があれば、ステップ 804 の処理へ戻る（ステップ 805）。

【0060】

コピー処理が完了してれば、記憶装置 110 は、コピー処理の完了を、コマンド処理 132 へ報告し、処理を終了する（ステップ 806）。

【0061】

本実施形態により、ホスト 100 はコピー処理の指示だけを行えば良く、実際のデータ転送は記憶装置が行うので、ホストやネットワーク等にかかる負荷が軽減されることとなる。

【0062】

図 9 は、本発明の第 2 実施形態が適用された計算機システムの構成図である。本実施形態では、データ転送領域情報 145 に含まれる同期状態 504 及び差分ビットマップが使用される点及びコマンドボリューム 900 を有する点が、第 1 実施形態と異なる。以下、第 2 実施形態に特有な部分のみを説明する。

【0063】

同期状態 504 には、データ転送処理の転送先/元データ領域の同期ペア状態を示す情報が登録される。同期ペア状態の取りうる値としては、「ペア未形成」、「ペア形成中」、「ペア形成済」がある。「ペア形成中」の状態とは、転送が指示されたデータ領域間で、転送元から転送先へデータの転送処理が実行中である状態を示す。「ペア形成済」の状態とは、データ領域間でのコピー処理が完了し、同期ペアが形成されていることを示す。ただし、「ペア形成中」の状態において転送元データ領域のデータの更新があった場合、「ペア形成済」状態でも、同期ペアのデータ領域間に差分がある可能性がある。「ペア未形成」の状態とは、データ領域間のデータ転送が指示されていないか、データ転送が完了した後、ホスト 100 の指示により同期ペアが解除された状態であることを示す。ただし、この状態は、データ転送処理がそもそも存在しないか、完了してしまっている

状態なので、制御メモリ 1 1 2 上には、データ転送領域情報 1 4 5 は確保されない。したがって、実際に同期状態 5 0 4 に設定されるのは、「ペア形成中」と「ペア形成済」の 2 つである。

【 0 0 6 4 】

差分ビットマップ 5 0 5 は、ペア形成中およびペア形成済の状態において、コピー元データ領域内のデータ更新の有無を示す。情報量を削減するため、ディスク装置 1 5 0 の全てのデータ領域を、例えば 64KB の特定サイズの小領域に分割し、その小領域単位と差分ビットマップ 5 0 5 の 1 ビットを一対一に対応させる。差分ビットマップ 5 0 5 には、小領域に対するデータの更新の有無が記録される。

【 0 0 6 5 】

ディスクキャッシュの割り当てにおいても、同様にディスク装置 1 5 0 を小領域に分割し、小領域毎にキャッシュを割り当てることで、キャッシュ管理を簡単化することが多い。この場合、差分ビットマップ 5 0 5 の 1 ビットをキャッシュ割り当て単位である小領域 1 つもしくは複数に対応づけることで、ビットマップの設定を容易にできる。

【 0 0 6 6 】

コマンドボリューム 9 0 0 には、標準のプロトコルには含まれない特殊な処理要求（データ転送処理要求等）が、データとして書き込まれる。第 1 実施形態では、記憶装置 1 1 0 に対するデータ転送処理要求のコマンドとして専用コマンドを追加した。本実施形態では、通常のライト要求コマンドを用い、データ転送要求がデータとしてコマンドボリューム 9 0 0 に書き込まれることで、記憶装置 1 0 0 へデータ転送要求が発行される。

【 0 0 6 7 】

記憶制御プロセッサ 1 1 1 は、コマンドボリューム 9 0 0 に対するライト要求を受け付けると、ライトデータを処理要求として解析し、対応する処理を起動する。要求される処理をライト要求の延長で実行しても応答時間が問題にならない程度短いならば、記憶制御プロセッサ 1 1 1 は、要求処理を実行し、ライト要求と合わせて完了報告を行う。要求される処理の実行時間がある程度長い場合には

、記憶制御プロセッサ 111 は、一旦ライト処理について完了報告を行う。ホスト 100 は、周期的にその要求処理が完了したかをチェックする。

【0068】

次に、本実施形態での CPU 101 及び記憶制御プロセッサ 111 の動作を説明する。

【0069】

CPU 101 で動作するデータ再配置処理 131 及び記憶装置 110 の記憶制御プロセッサ 111 で動作するコマンド処理 132 が、協働して LV の再構成を行うのは、第 1 実施形態と同様である。

【0070】

図 10 は、本実施例のデータ再配置処理 131 のフロー図である。

【0071】

ステップ 1001、1002 は、図 6 のステップ 602、603 と同様の処理のため、説明は省略する。

【0072】

CPU 101 は、LV の転送対象領域に対応する全 PV 領域のデータ転送処理の要求を、記憶装置 110 へ一括して発行する。コマンドボリューム 900 へのライトデータに含まれる転送要求には、LV 対象領域に対応する全 PV 領域の範囲情報(各部分領域の位置情報(各 PV 番号、先頭アドレス、サイズ)のリスト)及び転送先 PV 領域の範囲情報等のパラメタが含まれる。記憶装置 110 へライト要求コマンドを発行したら、CPU 101 は、記憶装置 110 からの完了報告を待つ(ステップ 1003)。

【0073】

完了報告を受け取ったら、CPU 101 は、規定の時間が経過するのを待つ(ステップ 1004)。

【0074】

CPU 101 は、データ転送領域の同期ペア状態を参照する要求を記憶装置 110 へ発行し、その完了を待つ。同期ペア状態を参照するには、具体的には、CPU 101 は、同期ペア状態の準備要求を含むデータをコマンドボリューム 90

0へ書き込むためのライト要求を発行する。CPU101は、記憶装置110から完了報告を受けた後、コマンドボリューム900へのリード要求を発行する（ステップ1005）。

【0075】

CPU101は、得られた同期ペア状態が「ペア形成済み」であるか否かを判定する。同期ペア状態が「ペア形成済み」である場合、CPU101は、ステップ1007の処理を行う。同期ペア状態が「ペア形成済み」ではない場合、CPU101は、ステップ1004の処理に戻り、同期ペア状態が変更されるのを待つ（ステップ1006）。

【0076】

その後、ステップ601と同様に、CPU101は、LVをオフラインにする（ステップ1007）。

【0077】

CPU101は、データ転送の転送元領域と転送先領域で形成済みの同期ペアの解除要求を、コマンドボリューム900を用いて発行する（ステップ1008）。コマンドボリューム900へ同期ペア解除要求を転送するライト要求コマンドの完了報告が記憶装置110より報告されたら、CPU101は、ステップ1004と同じ方法で、データ領域の同期ペア状態を参照する（ステップ1009）。

【0078】

取得された同期ペア状態が「ペア未形成」でなければ（ステップ1010）、CPU101は、規定時間が経過するのを待ち（ステップ1011）、ステップ1009の処理に戻って同期状態504の内容を再取得する。同期状態504が「ペア未形成」となっていれば、CPU101は、ステップ1012以降の処理を行う。CPU101は、LVのLV-PV対応情報141を更新し、LVをオンラインにしてLV再配置処理を完了する。

【0079】

図11は、コマンド処理132のフロー図である。

【0080】

記憶装置 110 は、ホスト 100 から受け取ったコマンドの対象が、コマンドボリューム 900 であるか否かを判定する（ステップ 1101）。

【0081】

コマンドの対象がコマンドボリューム 900 であれば、記憶装置 110 は、コピー処理を起動し、その完了を待つ（ステップ 1102）。

【0082】

コマンドの対象がコマンドボリューム 900 でなければ、記憶装置 110 は、コマンドの種別を判定する。コマンドがリードかライトかでそれぞれステップ 1104、1110 へ遷移する（ステップ 1103）。

コマンドの種別がリードである場合には、記憶装置 110 は、図 7 のステップ 703 から 707 と同様のリード処理を行う（ステップ 1104～1108）。

【0083】

コマンド種別がライトである場合には、記憶装置 110 は、図 7 のステップ 709 から 712 と同様のライト処理を行う（ステップ 1110～1113）。

【0084】

ステップ 1114、1115 は、本実施形態に特有な部分で、データ転送処理中にホスト 100 の別の処理が転送中の LV のデータに対してアクセスする場合の処理部分である。

【0085】

記憶装置 110 は、ライトの対象となったデータに、データ転送領域として登録されているデータ領域が含まれていないかを判定する（ステップ 1114）。登録されているデータ領域が含まれている場合、記憶装置 110 は、データ転送領域のうち、更新された部分を特定し、更新部分に対応した差分ビットマップを設定する（ステップ 1115）。

【0086】

記憶装置 110 は、要求処理の完了をホスト 100 へ報告する（ステップ 1109）。

【0087】

図 12 は、コピー処理 133 のフロー図である。

【0088】

記憶装置110は、コマンドボリューム900へ送信されたコマンドの種別を切り分ける（ステップ1201）。

【0089】

コマンドの種別がライトの場合、記憶装置110は、コマンドボリューム900へのライトデータの内容を解析し、処理要求内容や、転送元/先データ領域範囲指定の妥当性を判定する（ステップ1202）。妥当性に問題のある場合は、記憶装置110は、エラーを上位処理へ報告して処理を中断する（ステップ1203）。

問題ない場合は正常終了を報告し、記憶装置110は、ライトデータとして送られてきた処理要求の種別を判定する（ステップ1204）。

【0090】

処理要求がデータ転送の場合、記憶装置110は、図8のステップ802から805と同様のデータ転送処理を行う（ステップ1205～1208）。ただし、ステップ1205におけるデータ転送領域情報145の初期化処理では、同期状態504を「ペア形成中」にし、差分ビットマップ505を全て0クリアする処理を行う。データ転送処理が完了したら、記憶装置110は、同期状態504を「ペア形成済み」に変更し、処理を終了する（ステップ1209）。

【0091】

処理要求がペア解除の場合、記憶装置110は、解除対象のデータ転送領域ペアのデータ転送領域情報145を参照し、差分ビットマップ505に0nが設定されている部分がないかをチェックする（ステップ1210）。差分ビットマップ505に0であるビットが存在する、すなわちデータ転送領域に転送元/先で同期していない部分がある場合、未反映データを転送先データ領域へ転送する（ステップ1212）。記憶装置110は、ステップ1210へ戻り、差分ビットマップ505のチェックをやり直す。

【0092】

転送元/先のデータ転送領域で同期が取れたら、記憶装置110は、データ領域のデータ転送領域情報145をクリアし、データ転送領域情報145を格納す

る制御メモリ 112 の領域の割り当てを解除し、処理を終了する（ステップ 1213）。

【0093】

処理要求がペア状態準備である場合には、記憶装置 110 は、要求されたデータ転送領域の同期ペア状態を準備する（ステップ 1214）。データ転送領域がまだ存在し、制御メモリ 112 上にメモリ領域が割り当てられていれば、同期ペア状態として、同期状態 504 の値を用いる。データ転送領域が存在せず、メモリ領域が割り当てられていなければ、同期ペア状態として、「ペア未形成」を用いる。ステップ 1214 で準備された同期ペア状態は、コマンドボリューム 900 に対して発行されたリード要求のリードデータとして転送される（ステップ 1215、1216）。

【0094】

本実施例に拠れば、第一実施例よりも LV をオフラインにする時間を短くした上で LV の再配置をホスト等に負担を掛けないで行うことが出来る。

【0095】

第 3 実施形態について説明する。

【0096】

第 3 実施形態のシステム構成は、基本的に第 1 及び第 2 の実施形態と同一である。ただし、本実施形態では、記憶装置 110 が、ホスト 100 で使用されていないディスク装置 150 の記憶領域を管理する。そして、記憶装置 110 が、転送先の PV として使用されるデータ領域の条件、例えば、領域長、領域が格納される論理ユニット番号、接続に用いる接続ポート 114 の番号およびディスクタイプなどの指示をユーザ又は保守員から受ける。記憶装置 110 は、ユーザの指示に合致するデータ領域をホスト 100 が使用していないデータ領域から選びだし、ユーザ又は保守員に提示する。この点で、第 1、第 2 実施形態と異なる。以下、記憶装置 110 における情報の提示の方法について説明する。

【0097】

記憶装置 110 は、ホスト 100 が使用していないディスク装置 150 の番号のリストを、未使用領域管理情報として保持・管理する。

【 0 0 9 8 】

記憶装置 1 1 0 は、ユーザまたは保守員からの領域確保の指示を受け取ると、以下の処理を行う。

【 0 0 9 9 】

領域確保の指示にしたがい、記憶装置 1 1 0 は、保持している未使用領域管理情報を検索して条件を満たす未使用ディスク装置 1 5 0 を選定する（ステップ 1 - 1 ）。

【 0 1 0 0 】

記憶装置 1 1 0 は、選定したディスク装置 1 5 0 番号をユーザまたは保守員へ報告する（ステップ 1 - 2 ）。

【 0 1 0 1 】

ユーザ又は保守員は、記憶装置 1 1 0 から未使用のディスク装置 1 5 0 の番号を得ると、以下の手順にしたがって、データ転送の指示を行う。

ユーザ又は保守員は、未使用のディスク装置 1 5 0 の OS 管理情報を設定する。たとえば、UNIX OS では、未使用のディスク装置 1 5 0 に対して、デバイスファイル名を定義する（ステップ 2 - 1 ）。

【 0 1 0 2 】

ユーザ又は保守員は、OS 管理情報が設定されたディスク装置 1 5 0 について、LVM が使用することができるように、PV として定義する（ステップ 2 - 2 ）。

【 0 1 0 3 】

ユーザ又は保守員は、新しく定義した PV を転送先として指定して、記憶装置 1 1 0 に対して本発明によるデータ転送を指示する（ステップ 2 - 3 ）。

【 0 1 0 4 】

データ転送が完了したら、ユーザ又は保守員は、LV - PV 対応情報 1 4 1 の更新を指示する（ステップ 2 - 4 ）。

【 0 1 0 5 】

本実施形態が適用されたシステムが、例えば RAID 装置などのように、複数のディスク装置 1 5 0 の記憶領域の一部又は全体の集合から構成する論理ディス

ク装置をホスト 100 へ提示するような記憶装置 110 である場合を考える。この場合、未使用領域管理情報は、各未使用領域が属するディスク装置 150 番号および先頭領域と領域長のリストで構成される。

【0106】

記憶装置 110 は、次の手順で未使用領域を確保する。

【0107】

記憶装置 110 は、領域確保の指示の条件に合ったディスク装置 150 の未使用領域を検索する（ステップ 3-1）。

【0108】

指示分の容量を確保できない場合は、記憶装置 110 は、確保できない旨をユーザに通知し、処理を終了する（ステップ 3-2）。

【0109】

指示条件を満たすディスク装置 150 を一つ見つけたら、記憶装置 110 は、ディスク装置 150 に属する未使用領域の容量を確認し、指示分の領域があれば、指示分の領域を確保する。

具体的には、記憶装置 110 は、未使用領域管理情報から確保領域の登録を解除する。ディスク装置 150 の未使用領域容量が指示された分に満たない場合、未使用領域全体を確保し、別のディスク装置 150 を検索し、条件に合った未使用領域を確保する（ステップ 3-3）。

【0110】

記憶装置 110 は、ステップ 3-3 の処理を、指示分の容量のデータ領域を確保するまで繰り返す（ステップ 3-4）。

【0111】

記憶装置 110 は、該記憶装置が有する論理物理変換テーブルへ、確保された領域を登録することで、確保された領域で構成される論理ディスク装置を定義する（ステップ 3-5）。

【0112】

記憶装置 110 は、定義した論理ディスク装置の情報を使用者等に報告する（ステップ 3-6）。

【0113】

記憶装置 110 への未使用領域確保の指示は、第1の実施形態と同じく専用のコマンドを用いてもよいし、第2の実施形態と同じくコマンドボリュームへの命令書き込みでも構わない。保守用に記憶装置 110 へ接続するサービスプロセッサを設け、サービスプロセッサからコマンドを発行する構成も考えられる。

【0114】

ステップ 1-1 からステップ 3-4 までの一連の処理をスクリプト化することも可能である。その場合、ユーザまたは保守員は、転送先データ領域の選定条件をより詳細に指定することで、転送先データ領域を選定し、データ転送を自動的に実行することが可能となる。転送先データ領域の条件としては、記憶装置 110 の記憶領域での連続性、格納されるディスク装置 150 の物理容量、ヘッド位置付け時間やデータ転送速度などのアクセス性能が考えられる。RAID 装置の場合には、RAID レベルなどの物理構成条件なども転送先データ領域の条件として考えられる。特定の LV と物理構成、すなわちディスク装置 150 やディスク装置 150 が接続される内部バスおよび記憶制御プロセッサ 111 を共有しないという条件も考えられる。

【0115】

第3実施形態の変形として、転送先領域を指定せずに、転送元 LV と転送先領域の選定条件のみをユーザが指定することで、データ転送を指示する方式も考えられる。この場合、記憶装置 110 は、転送先領域を選定条件に従って選定し、新規に生成した論理ディスク装置へデータの転送を行う。記憶装置 110 は、データの転送が完了したら、転送の完了及び転送先として選定した領域情報をホスト 100 へ報告する。報告を受けたホスト 100 は、報告された転送先論理ディスク装置に対して、ステップ 2-1、2-2 及び 2-4 の処理により LV の移動を完了する。このときステップ 2-2 で、論理ディスク装置内のデータは有効なまま PV が定義される必要がある。

【0116】

なお、本発明は上記の実施形態に限定されず、その要旨の範囲内で数々の変形が可能である。

【0117】

各実施形態では、P V、ホスト 100 に提供される論理ディスク装置は、実際のディスク装置 150 と一対対応であるとしているが、P Vは、記憶装置 110 内で R A I D 5 レベルなどの R A I D で構成されていてもよい。その場合、ホスト 100 は、記憶装置 110 が提供する論理ディスク装置へ I / O を発行する。記憶制御プロセッサ 111 は、論理ディスク装置への I / O を、論理物理変換によりディスク装置 150 への I / O に変換する。

【0118】

各実施形態では、データの再配置の例として、L V M の管理する L V の再配置を採用しているが、L V が未割り当てな P E を連続化（ガベージコレクション）する処理や、D B M S の管理する D B 表の再配置の処理など、他のデータ再配置にも本発明を適用することが可能である。

【0119】

また、ホスト 100 と記憶装置 110 間でのデータ転送処理要求方法について、第 1 の実施形態の専用コマンドを使用する方法と、第 2 の実施形態のコマンドボリューム 900 を利用する方法は、両実施形態で入れ替えても実現可能である。

【0120】

各実施形態では、転送先 P V が 1 台の場合を想定しているが、複数台としても構わない。その場合には、複数台の P V のそれぞれについて、転送元データをどのように分散させるかを指定する必要がある。複数の P V への分散の方法としては、各 P V に均等に分配する場合、P V の指定順に空き容量の許す限り詰め込んでいく場合が考えられる。また、均等に分配する場合には、さらに、各 P V 内でデータを連続させる場合、特定サイズでデータが分割され、R A I D におけるストライピングの如く分割されたデータを各 P V に順に格納する場合などが考えられる。

【0121】

ユーザ又は保守員があらかじめ転送先として連続領域を P E に割り当ててから、データ再配置処理 131 にパラメタとして引き渡す、あるいは、データ再配置

処理内で前述した空き P E のガベージコレクションを行っても良い。

【0122】

各実施形態では、転送先データ領域へのアクセスは発生しないことを前提としている。つまり、記憶装置 110 での転送先データ領域に対するアクセス抑止の考慮はなく、転送先データ領域へのアクセスがあると、そのままアクセス対象となる領域に対してデータ参照/更新が行われる。しかし、ホスト 100 でのアクセス抑止の保証がないケースにそなえて、記憶装置 110 で、データ転送領域の転送先データ領域として登録されているデータ領域に対する I/O を拒否する構成も考えられる。逆に、データ転送処理の完了を待たずに、ホスト 100 が、L V - P V 対応情報 145 を再配置後の状態に更新し、転送対象の L V へのアクセスを、転送先 P V で受け付ける方法をとってもよい。この場合、記憶装置 110 でデータ転送領域のペアを造り、同期化のためのデータコピーを行うのは第 2 の実施形態と同じである。ただし、転送先領域へのリード要求に対しては転送元領域のデータに参照し、転送先領域へのライト要求に対しては、転送元領域にもデータの反映を行う必要がある。

【0123】

本実施形態によれば、通常よりも短い時間、L V をオフラインとするだけで L V の再配置処理をおこなうことができ、システムの可用性を向上することができる。

【0124】

【発明の効果】

本発明の計算機システムによれば、記憶装置に格納したデータを別の領域に移動する際に、データ転送処理を記憶装置内で行うことで、ホストおよびチャネルの負荷を削減することが可能となる。

【0125】

また、本発明の計算機システムによれば、データ再配置におけるデータ転送中に当該データに対するアクセスを受けつけることが可能となる。この結果、データ再配置における対象データをアクセスする業務の停止時間を短縮することが可能となる。

【図面の簡単な説明】**【図 1】**

本発明の第 1 の実施形態が対象とする計算機システムのブロック図である。

【図 2】

本発明が想定しているデータ再配置処理の概要図である。

【図 3】

本発明における L V 管理情報の構成図である。

【図 4】

本発明における P V 管理情報の構成図である。

【図 5】

本発明におけるデータ転送領域情報の構成図である。

【図 6】

本発明の第 1 の実施形態におけるデータ再配置処理のフロー図である。

【図 7】

本発明の第 1 の実施形態におけるコマンド処理のフロー図である。

【図 8】

本発明の第 1 の実施形態におけるコピー処理のフロー図である。

【図 9】

本発明の第 2 の実施形態における計算機システムの構成図である。

【図 1 0】

本発明の第 2 の実施形態におけるデータ再配置処理のフロー図である。

【図 1 1】

本発明の第 2 の実施形態におけるコマンド処理のフロー図である。

【図 1 2】

本発明の第 2 の実施形態におけるコピー処理のフロー図である。

【符号の説明】

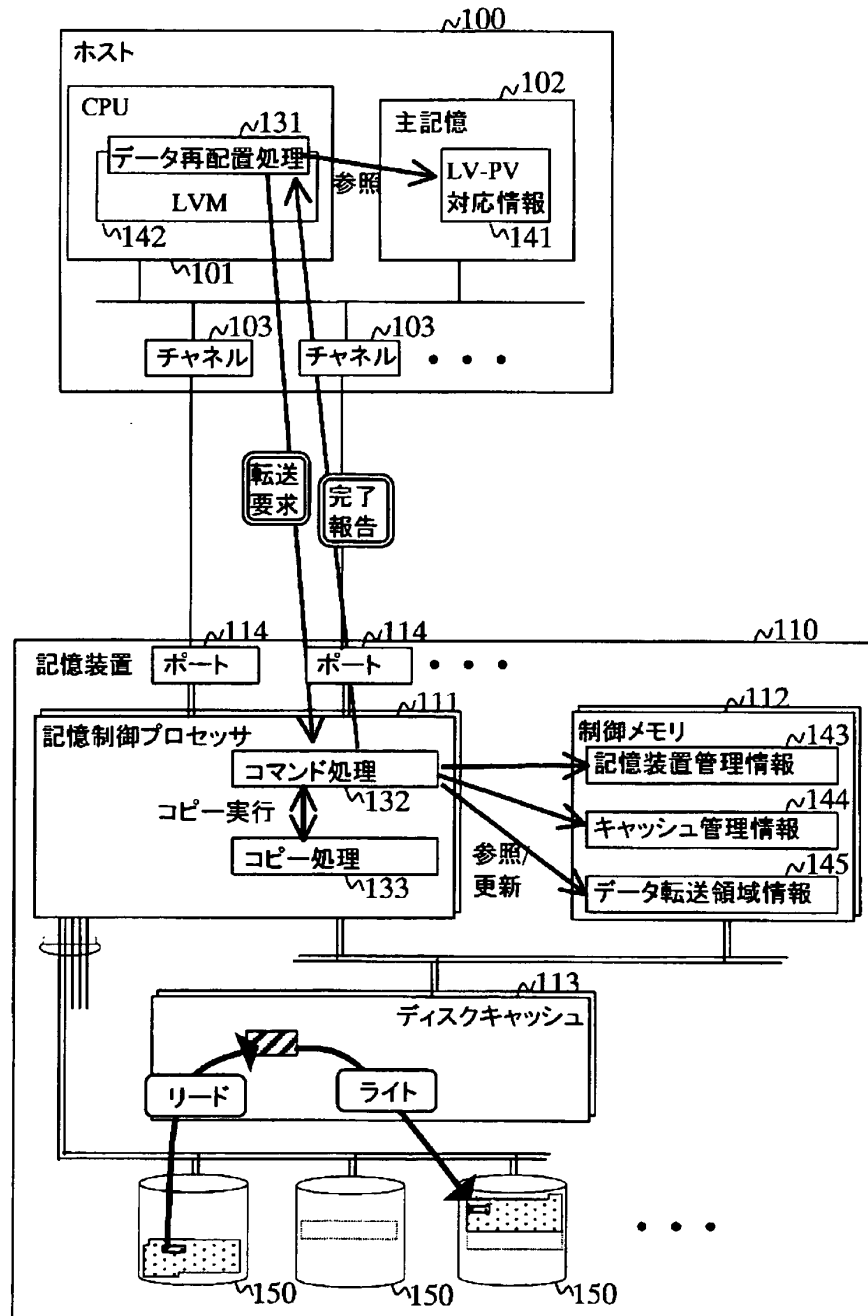
1 0 0…ホスト、1 0 1…C P U、1 0 2…主記憶、1 0 3…チャネル、1 1 0…記憶装置、1 1 1…記憶制御プロセッサ、1 1 2…制御メモリ、1 1 3…ディスクキャッシュ、1 4 1…L V - P V 対応情報、1 4 3…記憶制御管理情報、1

4 4 …キャッシュ管理情報、1 4 5 …データ転送領域情報、1 5 0 …ディスク装
置、9 0 0 …コマンドボリューム

【書類名】 図面

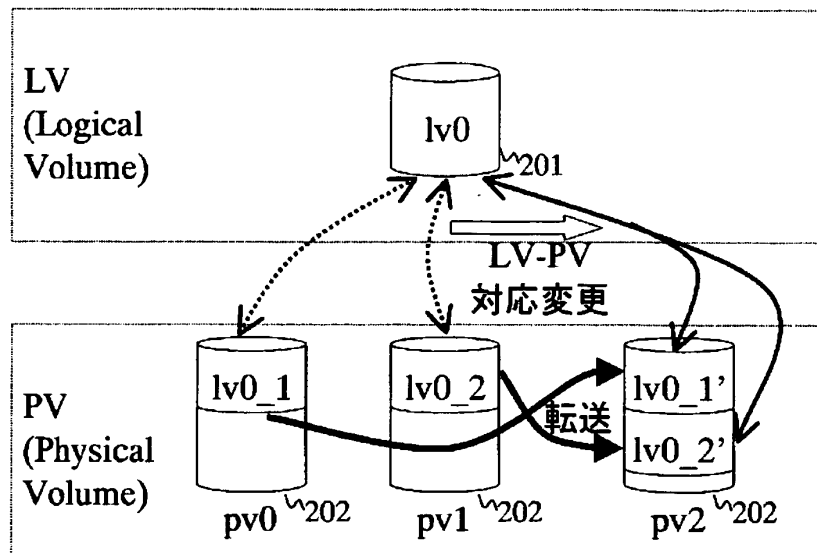
【図 1】

図 1



【図 2】

図2



【図3】

図3

LV管理情報300

PVリスト	~301
LE数	~302
LEサイズ	~303
LE-PE対応情報	~310

LE-PE対応情報310

~311 LE番号	~312 PV名	~313 PE番号
0	/dev/dsk/c0t0d0	500
1	/dev/dsk/c0t0d0	501
2	/dev/dsk/c0t0d0	502
3	/dev/dsk/c0t0d0	503
	• • •	
499	/dev/dsk/c0t0d0	999
500	/dev/dsk/c1t0d0	100
501	/dev/dsk/c1t0d0	101
502	/dev/dsk/c1t0d0	102
	• • •	

【図 4】

図4

PV管理情報400

LVリスト	~401
PE数	~402
PEサイズ	~403
PE-LE対応情報	~410

PE-LE対応情報410

~411 PE番号	~412 LV名	~413 LE番号
0	/dev/vg0/lv other	0
1	/dev/vg0/lv other	1
2	/dev/vg0/lv other	2
3	/dev/vg0/lv other	3
	•	
	•	
	•	
499	/dev/vg0/lv other	499
500	/dev/vg0/lv0	0
501	/dev/vg0/lv0	1
502	/dev/vg0/lv0	2
	•	
	•	
	•	

【図 5】

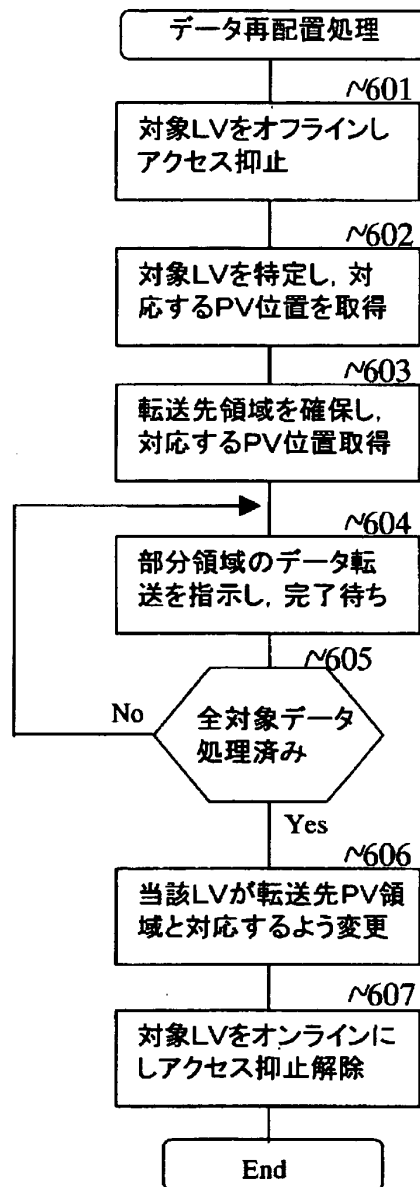
図5

データ転送領域情報145

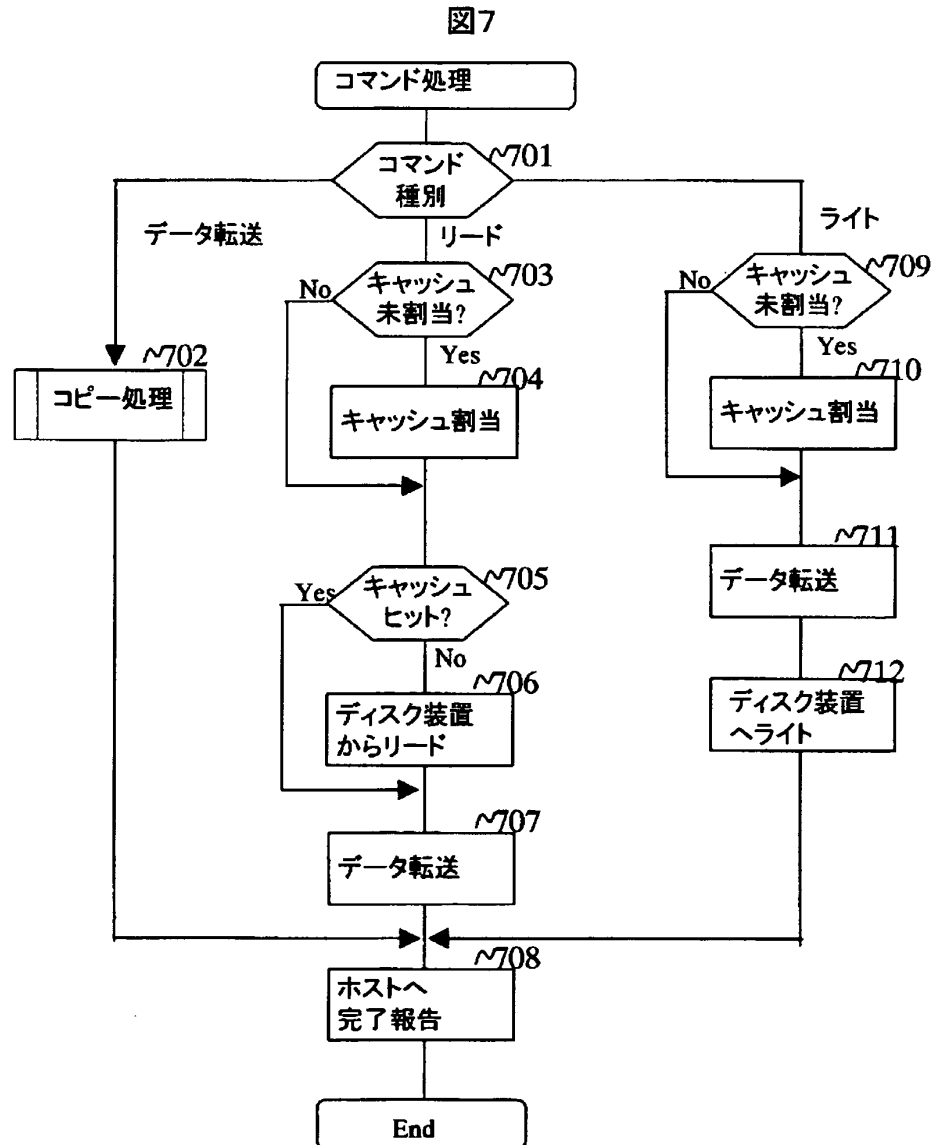
転送元範囲情報	~501
転送先範囲情報	~502
進捗ポインタ	~503
同期状態	~504
差分ビットマップ	~505

【図 6】

図6

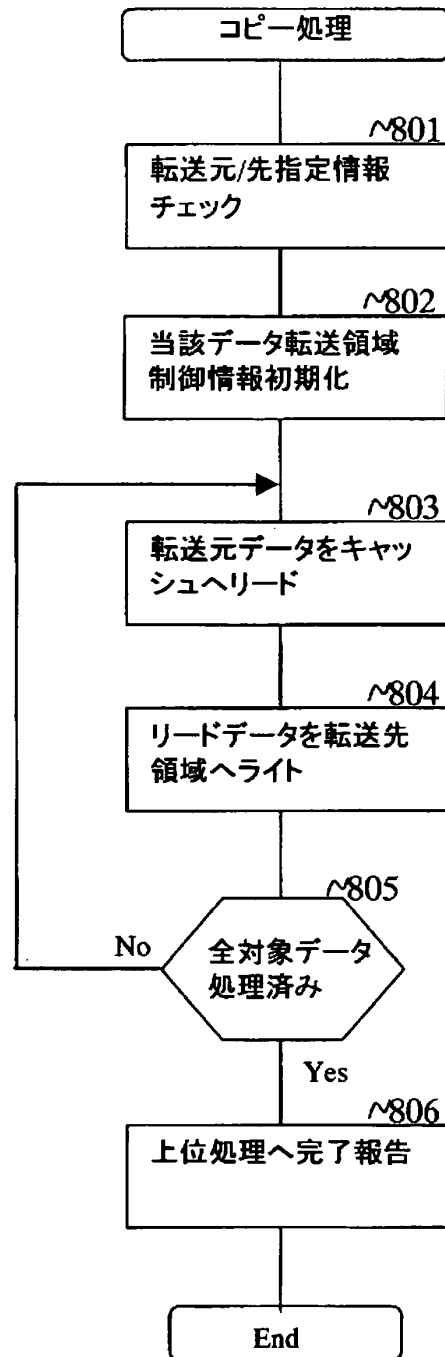


【図 7】

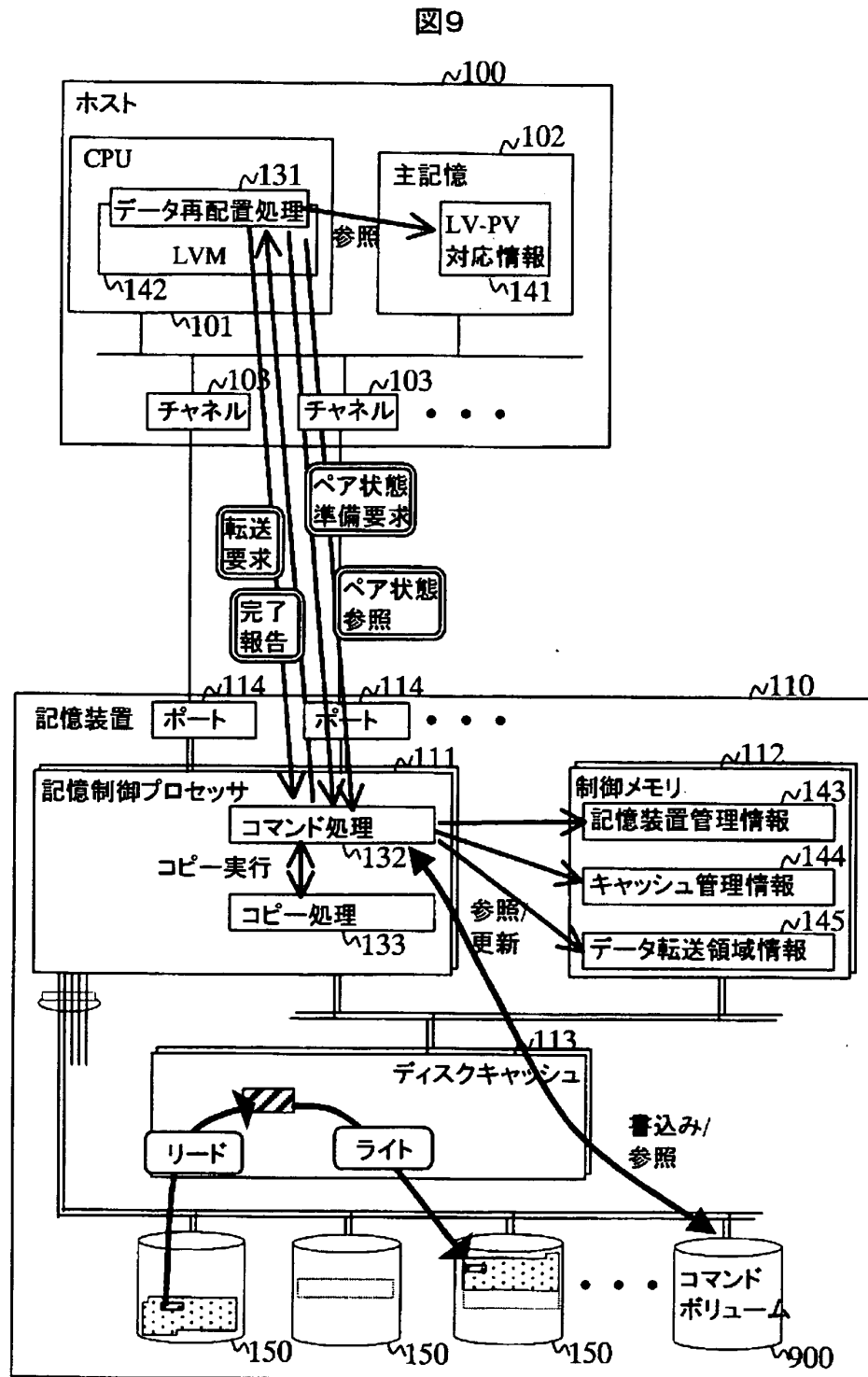


【図 8】

図 8

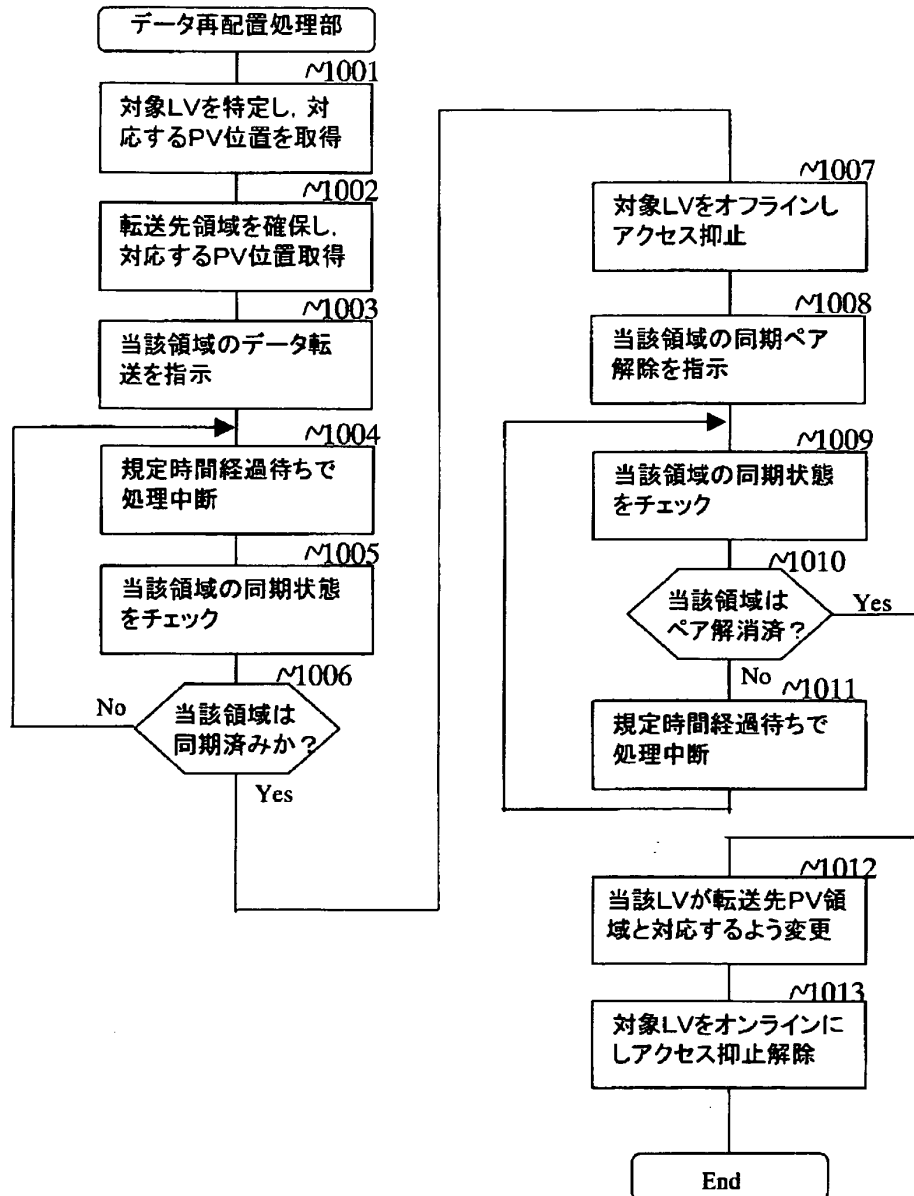


【図9】

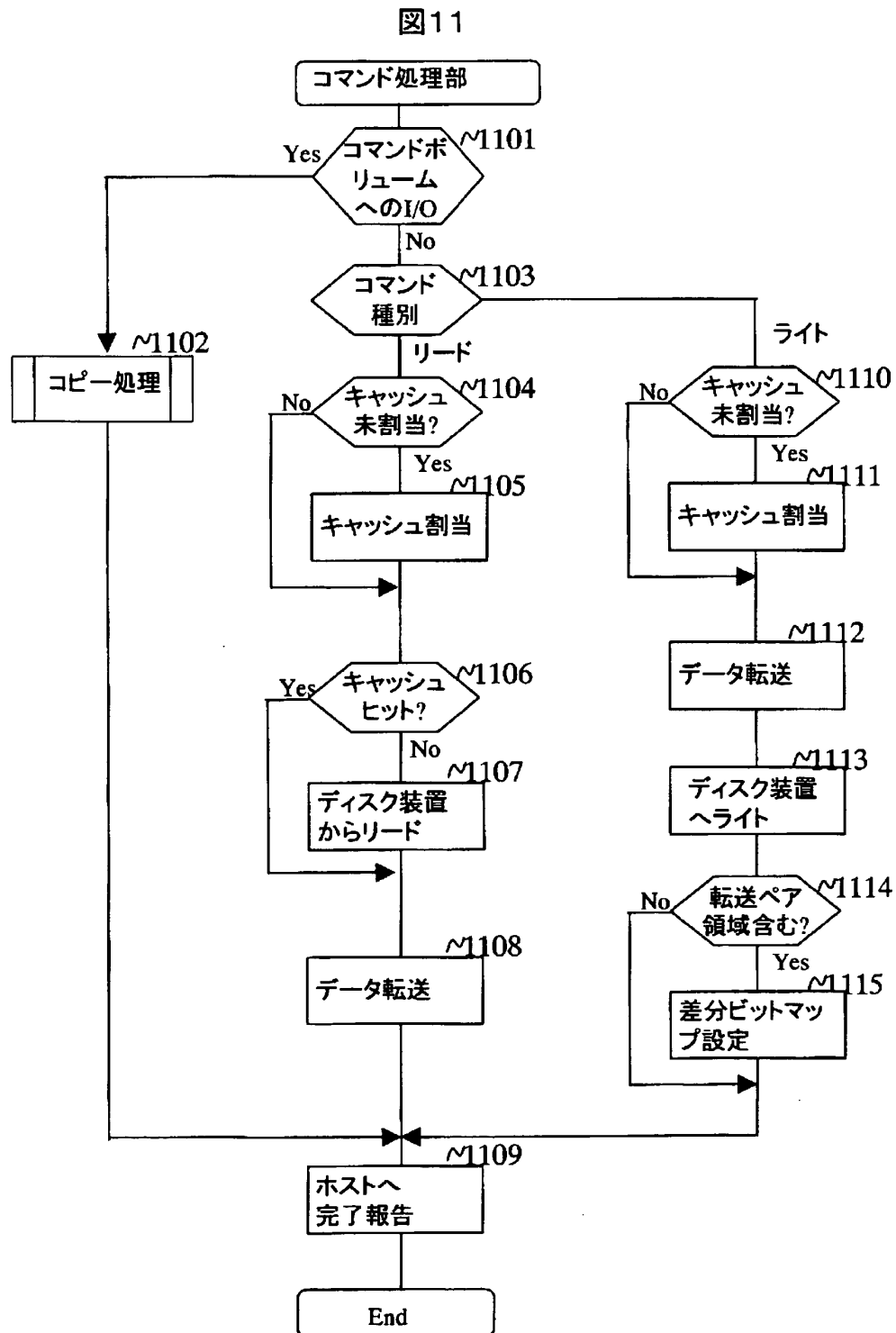


【図10】

図10

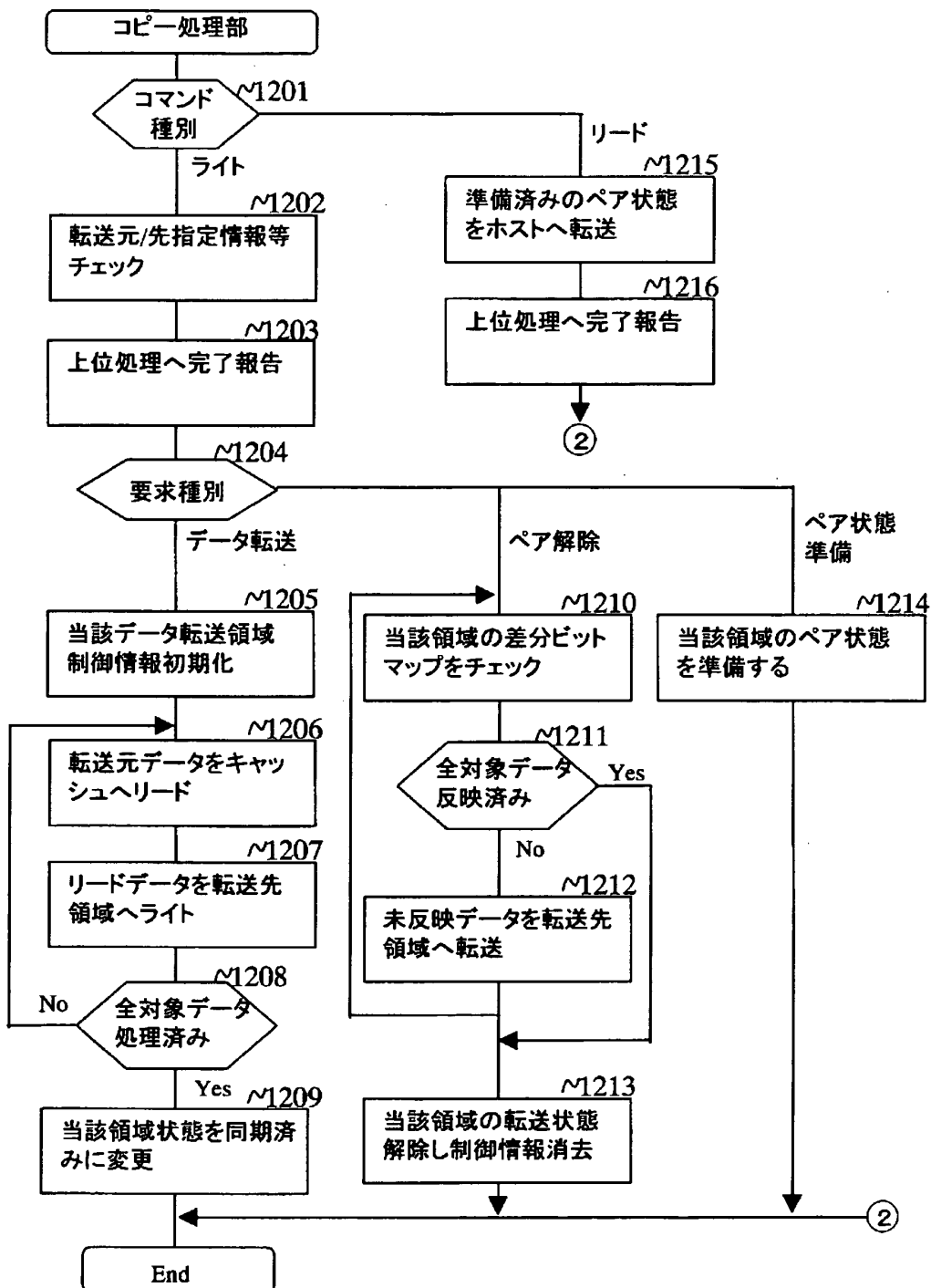


【図11】



【図12】

図12



【書類名】 要約書

【要約】

【課題】

ストレージに格納したデータを別の領域に再配置する処理において、領域間のデータ転送をホストで行うため、ホストおよびチャネルに負荷がかかる。

【解決手段】

ホストはデータ転送元/先の領域情報をパラメタとしてストレージへデータ転送を依頼する。ストレージは内部で転送元のディスク装置から転送先のディスク装置へデータ転送を行う。対象データの転送が全て完了したら、ホストはそのデータの格納場所を転送先領域に変更して登録する。

【選択図】 図 1

特願 2001-053472

ページ: 1/E

出願人履歴情報

識別番号

[000005108]

1. 変更年月日

1990年 8月31日

[変更理由]

新規登録

住所

東京都千代田区神田駿河台4丁目6番地

氏名

株式会社日立製作所